DOCUMENT RESUME

ED 064 349 TM 001 516

AUTHOR Harris, Chester W.
TITLE An Index of Efficiency for Fixed-Length Mastery Tests.
PUB DATE Apr 72
NOTE 10p.; Paper presented at the annual meeting of the American Educational Research Association (Chicago, Illinois, April 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Models; *Statistical Analysis; *Student Distribution; *Student Testing; *Test Reliability

ABSTRACT

The efficiency of mastery tests of fixed length which sorts students into two categories is analyzed. For the sort of the students, an index, suggested by Fisher's linear discriminant function for two groups, is provided. (DB)

ED 064349

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

AN INDEX OF EFFICIENCY FOR FIXED-LENGTH MASTERY TESTS

Chester W. Harris

University of California, Santa Barbara

TM 001 516

AERA

Chicago, Illinois

April 4, 1972

First, let us limit the territory to be explored. We are concerned here only with fixed length mastery tests. Kriewall (1969) has considered the adaptation of an acceptance sampling plan to mastery testing in mathematics; his procedure fixes Type I and Type II error probabilities but allows the number of items to vary, and thus is appropriate for variable length mastery tests. In passing I would like to remark that Livingston's (1971) comment that Kriewall's procedure assumes that the test items are homogeneous in difficulty for each examinee seems to be in error. For such a sequential test, Wald (1947, p. 88) defines a parameter p: "Let p devote the unknown proportion of defectives in the lot." He then proceeds to define a random variable which takes values of 0 and 1 and develops a sequential test of

$$H_0: \ p \leq p'$$
$$H_1: \ p > p'$$

where p' is a chosen error rate. For Wald p characterizes a lot (a possibly finite population) and in no sense characterizes <u>each</u> unit being inspected, or each item, as Livingston implies. Within Wald's scheme, to speak of the units being homogeneous in diffuclty seems to be meaningless; each unit is either defective or not as determined by the particular inspector. All of this translates readily into a procedure for testing an hypothesis about the proportion of items in a lot a given student can pass. I offer this point to Livingston and Kriewall for further discussion.

I shall assume that fixed length mastery tests are intended to be used in connection with a program of instruction. The interest is, through appropriate instruction, to bring one or more students to a mastery of a limited and reasonably specific instructional objective or class of objectives in a reasonable period of time. For any such objective or class of objectives the expectation is that a satisfactory (valid) mastery test will function to sort students into two groups: those who have and those who have not "mastered" the objective or class of objectives.

Let us ask how we might assess the validity of such a mastery test and in the process identify two areas or problems that we merely mention and then leave unexplored. We would like to be able to develop for a sample of students the fourfold table that would result from classifying them as "true masters" or "true nonmasters" based on criterion data, and as "indicated masters" or "indicated nonmasters" based on the mastery test. Given such a fourfold table for a sample of students, we would also like to choose a statistic to summarize the strength of this observed relation and possibly estimate its strength in the population. There obviously are several such statistics from which we might choose, and equally obviously such a table presents an occasion for arguments concerning the appropriateness and the merits of the several options. But arguments about appropriateness would turn at least in part on conceptions of and an examination of the methods used to gather criterion data and the methods used to gather the mastery test data, and so the bases of these arguments would be laid before the actual choice of an index was made. I wish to set this problem of the choice of a statistic

aside even though I recognize that what I propose with respect to gathering mastery test data bears on it.

I also wish to set aside without adequate discussion questions concerning the nature of the relevant criterion data and methods of gathering them. For example, one is prompted to ask whether or not latent trait theory gives an appropriate conception of a mastery criterion, and if it does whether the relevant observations should be univariate or possibly multivariate with a specified underlying covariance structure. In contrast one might ask whether or not a transfer task is the appropriate model for a mastery criterion, and if it is, what principles are available for selecting, administering, and scoring such a task. Still another question concerns the relevance of the experimental history of the student as a criterion datum. Perhaps the notion of mastery testing is moving us in the direction of considering tutored and nontutored groups as criterion groups, and will require that we develop a more explicit role in validity studies for the data provided by the experimental or instructional history of the student. If, as seems reasonable, mastery testing is to be validated for selection purposes--i.e., selection for further instruction within an instruction network--then the ultimate criterion may be the student's advancing through this network or hierarchy to an end point.

Let us now turn to the question of what I shall call the efficiency of a mastery test. I am aware that this term is used in connection with classic reliability theory, but at the outset I do not intend to imply that the term <u>efficiency</u> means all that reliability means.

A necessary characteristic of a mastery test is that it sorts students into two categories. If in addition the test is valid, it will tend to sort them into the correct two categories: that is, into the categories determined by the criterion data. In the absence of such criterion data, it may be informative simply to examine how well the test sorts defined samples of students into categories and possibly to measure its efficiency in this sense. It is important to point out that we are not breaking a new path here, since as early as 1936 Richardson (1936) considered this problem of a "criterion of two categories" using scores on a total test cut at various points as the "criterion." His work relates the difficulty of a test element to the prediction of a two-category criterion, employing certain distributional assumptions. We shall attempt a similar development making somewhat different assumptions.

Let us assume that a mastery test consists of K items and that a total score on the test is derived by summing the number of correct items, which gives 0 and K as the limits for any score. Let us also think of these items as ones for which the student produces a response, rather than chooses a given alternative. With total scores ranging possibly from 0 through K, there are K different possible separations into two groups on the basis of total test score. For example, students who score K may be sorted into one group and all others into the other group; students who score at least K-1 may be sorted into one group and all others into the other goup; etc. Thus, there can be K different sorts. For any sort, let us develop an index that is suggested by Fisher's linear discriminant function for two groups (Fisher, 1936). The discussion by Tatsuoka (1971, chapter 6) is quite helpful since he shows canonical correlation equivalents of discriminant functions.

By a "sort" we mean that the sample has been sorted into two groups on the basis of some cutting point on the total score for the K items. We can then, following Fisher, develop two K by K matrices, B and W. (See Tatsuoka, p. 158-159). The matrix W is the pooled within groups sum of squares and cross products of the item responses. The matrix B equals T-W, where T is the sum of squares and cross products of the item responses, ignoring the separation into two groups. Then given the group membership, the Fisher discriminant function is

$$\frac{v' \; B \; v}{v' \; W \; v} = \lambda,$$

where v is a column vector of weights, chosen to maximize λ. Instead of using these weights, let us use an a priori vector of equal weights, 1, and form the function

$$\frac{1' \; B \; 1}{1' \; W \; 1} = \lambda_c$$

which is a special case (equal weights) corresponding to using the total score (sum of the item scores) to discriminate the two groups. Generally λ_c is less than λ.

Now λ_c turns out to be a function of the sums of squares associated with the two group analysis of variance. It is

$$\lambda_c = \frac{SS_b}{SS_w},$$

where SS_b and SS_w refer to the analysis of the total scores on the K items for the two groups. We also know that in general the Fisher discriminant

function can be related to a canonical correlation between the given variables (items) and a dummy variable indicating group membership. In general, if μ^2 is the squared canonical correlation, then

$$\lambda = \frac{\mu^2}{1 - \mu^2} \quad ,$$

and

$$\mu^2 = \frac{\lambda}{1 + \lambda} \quad .$$

An analogous treatment of λ_c yields

$$\mu_c^2 = \frac{\lambda_c}{1 + \lambda_c}$$

or

$$\mu_c^2 = \frac{SS_b}{SS_b + SS_w} \quad .$$

This coefficient is equivalent to the squared Pearson product-moment correlation between total score on the test and the dummy variable designating the sort. Thus it is the squared point biserial correlation coefficient.

Sorting into two (non empty) groups on the basis of total test score necessarily yields a positive valve for SS_b and thus a positive valve for μ_c^2. The upper limit of μ_c^2 can be +1 when $SS_w = 0$; this could occur, for example, when only two different total scores appeared in the sample and we sorted into the obvious two groups. Such a situation would correspond to a perfect <u>phi</u> coefficient and thus is not in conflict with the well-known principle that the point biserial cannot take values of $\pm$ 1.

The coefficient μ_c^2 for a given sort based on total score measures the extent to which the sum of the K item scores (0, 1 scores) can discriminate the two groups defined by the sort. It is a measure of efficiency in this sense and has two features that make it an analog of a classic reliability

coefficient. One is that it can be conceived as the ratio of true score variance to observed score variance for a particular definition of true score. To achieve this correspondence, assign to each individual in the upper group a true score equal to the mean of the upper group and to each individual in the lower group a true score equal to the mean of that lower group. Then μ_c^2 will be the ratio of the variance of these assigned true scores to the variance of the observed scores. Note that μ_c^2 was defined originally without reference to true score variance and that we have now simply answered the question of how true scores might be conceived to make μ_c^2 an analog of the classic reliability coefficient.

The second feature is that the largest μ_c^2 for a given test is an upper limit to the validity of the mastery test when validity is measured in an analogous fashion. First note that for a K item test there are K different sorts into two groups based on total score and that there is a value of μ_c^2 associated with each sort. Suppose now we have a dichotomous criterion and use this, rather than total score, to sort students into two groups. If we now measure in a similar fashion the extent to which the sum of the item scores can discriminate the two criterion groups we find that this coefficient cannot exceed the largest μ_c^2 . It may of course be substantially smaller. It also is true that if the two criterion groups are not equal in number, the upper limit will be some μ_c^2 less than the maximum and corresponding to a sort into two groups with the same relative frequencies.

It is possible to deduce some generalizations about maximum values of μ_c^2. For example, for symmetric distributions the maximum value of μ_c^2 occurs

when the proportion in the upper (or lower) group is close to one-half and decreases as this proportion diverges from one-half. For symmetric distributions of equal range, a rectangular distribution gives a larger maximum μ_c^2 than does a normal distribution. It is intuitively obvious that a U-shaped distribution has a larger maximum coefficient than does a rectangular distribution of the same range. Interestingly, a rectangular distribution of small range has a larger maximum coefficient than does a rectangular distribution of large range, though the difference may be small.

Our identification above of the manner in which true scores should be conceived in order to make μ_c^2 a ratio of true and observed score variances suggests that the corresponding item model is one in which items are uncorrelated within the upper group and also within the lower group. This is a "local independence" condition and corresponds to the notion that no latent trait that is a linear function of the item scores should distinguish among the individuals within a group (either upper or lower). If items are uncorrelated within both groups, then the population correlation between two items (equivalently, between an item and the sum of other items) for the combined populations is necessarily simply a function of differences in means for the two populations. This type of analysis leads to the notion of experimental induced correlation or "reliability" which others have been aware of.

This is perhaps enough discussion to indicate one means of dealing with mastery tests. The discussion may, hopefully, stimulate both theoretical and empirical work.

REFERENCES

Fisher, R. A., The use of multiple measurements in taxonomic problems. Annals of Eugenics, Volume VII, Part II, 1936, p. 179-188.

Kriewall, Thomas E., Applications of information theory and acceptance sampling principles to management of mathematics instruction. Doctoral Dissertation, University of Wisconsin, 1969.

Livingston, Samuel A., Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, p. 13.

Richardson, M. W., The relation between the difficulty and the differential validity of a test. Psychometrika, Volume I, Number 2, 1936, p. 33-49.

Tatsuoka, Maurice M., Multivariate analysis. New York, Wiley, 1971.

Wald, Abraham, Sequential analysis, New York. Wiley, 1947.